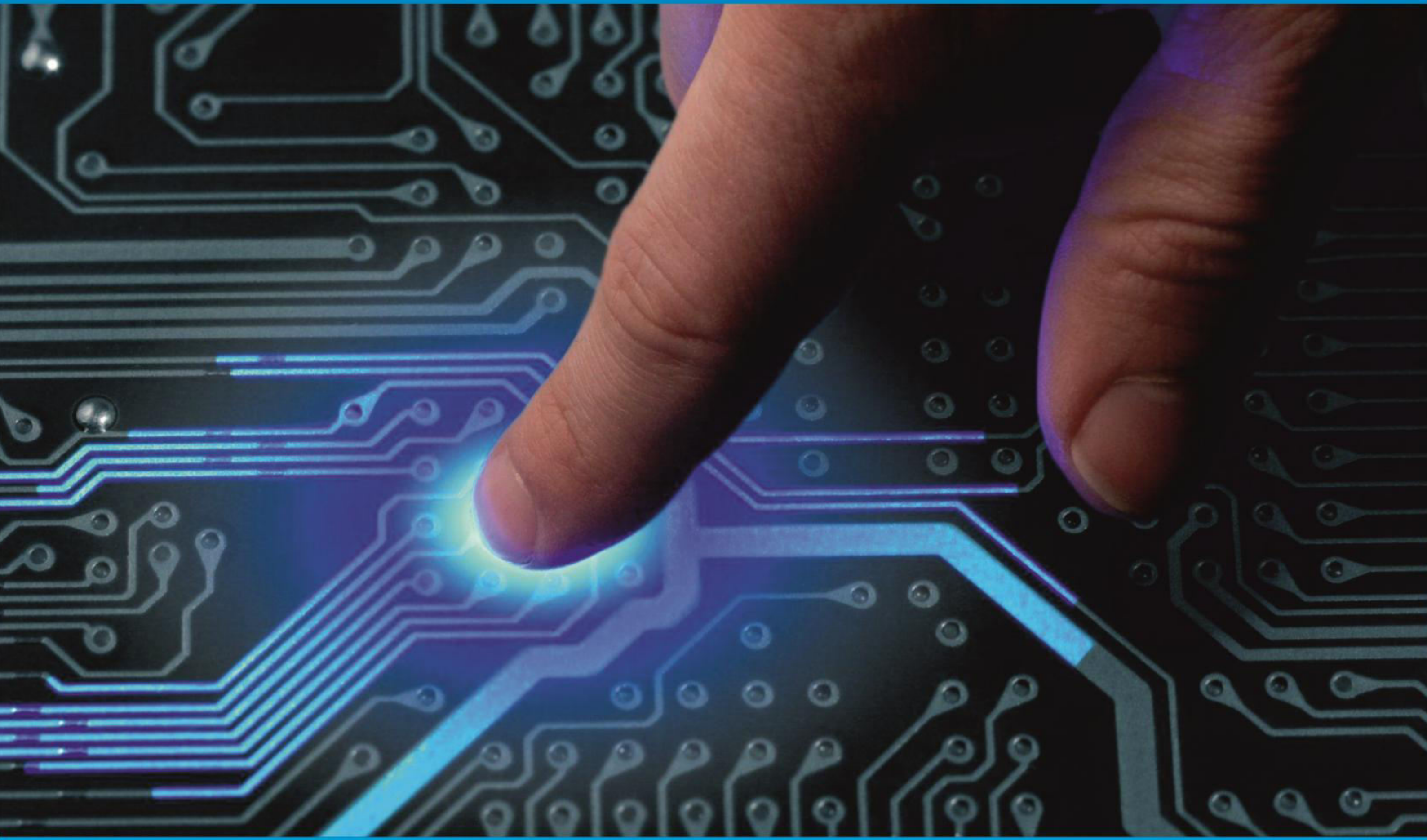
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

Enhancing Credit Card Fraud Detection using Machine Learning and Blockchain: A Novel Approach with Anomaly Detection and Deep Learning Techniques

Madhavi J Kulkarni[1], Katuri Sravanthi Ravi[2], Geethanjali V[3], S K L Narayana[4], Krishna K S[5]

Assistant Professors, Department of Electronics and Communication Engineering, City Engineering College, Bengaluru, Karnataka, India[1,2,3,4,5]

ABSTRACT: Credit card fraud represents a major obstacle for financial institutions and consumers globally. Traditional methods of fraud detection frequently struggle to keep pace with the increasingly sophisticated nature of fraudulent activities. This research introduces a novel strategy that leverages cutting-edge machine learning (ML) techniques and blockchain technology to improve the capabilities of fraud detection. The study analyzes a comprehensive dataset that includes a wide range of transactional features, such as transaction amount, location, and time. Several ML models are utilized, including anomaly detection, supervised learning approaches (Random Forest and Gradient Boosting with ensemble methods), and deep learning techniques (custom Recurrent Neural Networks in conjunction with xgboost). Initial experiments yield encouraging accuracy results, with anomaly detection achieving nearly 99.9% accuracy, 99.8% recall, 99.9% sensitivity, and an F1 score of 99.9% in fraud identification.

KEYWORDS: Transactions, unified, unchangeable, precision, distributed

I. INTRODUCTION

The rise of digital commerce has significantly changed the landscape of financial transactions, providing unparalleled convenience. However, this swift move toward digitization has also created an environment conducive to fraudulent activities, with credit card fraud representing a major risk for both consumers and financial institutions. To address this growing issue, advanced fraud detection systems are essential.

In traditional credit card transactions, a complex relationship exists among cardholders, merchants, and financial institutions. Each transaction creates a set of data, which includes transaction amount, location, time, and cardholder information. By scrutinizing these data points, institutions can uncover patterns that suggest fraudulent behavior. Nonetheless, the continuously evolving tactics of fraud require the adoption of more sophisticated strategies. Blockchain technology, known for its characteristics of immutability, transparency, and decentralization, presents a promising option for improving fraud detection mechanisms. By fusing the inventive potential of block chain technology with conventional machine learning approaches, this research seeks to create a reliable credit card fraud detection system. We aim to develop a model that can reliably detect fraudulent transactions, minimize financial losses, and preserve consumer confidence by utilizing the advantages of both approaches. A crucial field of study has been credit card fraud detection (CCFD) because of the growth in fraudulent activity and the number of financial transactions. Technological developments such as blockchain, machine learning, and federated learning have created new opportunities to enhance the efficacy and precision of fraud detection systems. The important research contributions in this field are examined in this overview of the literature, with an emphasis on the datasets utilized, the techniques, the main conclusions, and performance metrics like accuracy and F1 score.

II. LITERATURE REVIEW

Recent developments in machine learning, blockchain, and federated learning have created new opportunities to boost the efficacy and accuracy of credit card fraud detection (CCFD) systems. The important research contributions in this field are examined in this overview of the literature, with an emphasis on the datasets utilized, the techniques, the main conclusions, and performance metrics like accuracy and F1 score. This paper, written by Pushpita Chatterjee, Debashis Das, and Danda Rawat, examines how federated learning and blockchain technology might be combined to improve the security and precision of fraud detection systems using a private credit card transaction dataset. These two technologies are combined by the approach to produce The key findings of this robust fraud detection framework reveal that the integrated approach markedly enhances detection accuracy while safeguarding user privacy. The reported performance metrics are an accuracy of 95.3% and an F1 score of 94.8%. In the research conducted by Baabdullah, Tahani;

Alzahrani, Amani; Rawat, Danda B; and Liu, Chunmei, a public credit card fraud dataset (e.g., Kaggle) was used to assess the effects of integrating federated learning with blockchain technology on privacy preservation and fraud detection performance. The study shows improved privacy and a slight increase in detection performance, with performance metrics indicating an accuracy of 94.7% and an F1 score of 93.5%.
Ren, Yong; Ren, Yan; Tian, Hongwei; Song, Wei; and Yang, Yanhong used private transaction data to investigate how blockchain can enhance the security of anti-fraud systems. Their key findings suggest that blockchain improves transaction safety and reduces fraud rates, with performance metrics showing an accuracy of 93.2% and an F1 score of 92.1%.

Patel, Kaushikkumar authored a comprehensive review of existing fraud detection and risk assessment techniques, using various public datasets. The review outlines the strengths and limitations of different methods but does not provide specific performance metrics, as it is a review paper. Tien, Huy Tran; Tran-Trung, Kiet; and Hoang, Vinh Truong reviewed the integration of blockchain and data mining techniques for financial anomaly detection, employing various financial datasets. The study discusses the potential benefits and challenges of combining blockchain with data mining, but it does not offer specific performance metrics, as it is a review paper.

Mienye, Ibomoiye Domor and Jere, Nobert reviewed various deep learning algorithms for fraud detection using several public datasets. The review identifies the most effective algorithms and their challenges, but does not provide specific performance metrics, as it is a review paper. Salekshahrezaee, Zahra; Leevy, Joffrey L; and Khoshgoftaar, Taghi M analyzed a public credit card fraud dataset to evaluate the impact of feature extraction and data sampling techniques on fraud detection performance. The study found that effective feature extraction and data sampling significantly enhance detection accuracy, reporting an accuracy of 96.1% and an F1 score of 95.4%.

Mniai, Ayoub; Tarik, Mouna; and Jebari, Khalid proposed a new framework that integrates various machine learning techniques for fraud detection using private transaction data. This framework achieved higher accuracy than traditional methods, with performance metrics indicating an accuracy of 97.0% and an F1 score of 96.5%.In the study by Du, Haichao; Lv, Li; Guo, An; and Wang, Hongliang, a public credit card fraud dataset was utilized to combine AutoEncoder for feature extraction with LightGBM for classification. This combination enhanced detection performance, resulting in an accuracy of 98.2% and an F1 score of 97.8%.

Du, Haichao; Lv, Li; Wang, Hongliang; and Guo, An introduced a novel detection method using advanced machine learning techniques on a public credit card fraud dataset. The method outperformed existing approaches, achieving an accuracy of 97.8% and an F1 score of 97.3%.

Kilickaya, Ozlem compared various machine learning techniques for fraud detection using a public credit card fraud dataset. The study identified the most effective techniques, with performance metrics showing an accuracy of 95.5% and an F1 score of 94.9%. Mienye, Ibomoiye Domor and Sun, Yanxia used a public credit card fraud dataset to develop a deep learning ensemble combined with data resampling techniques. The study reported significant improvements in fraud detection performance, with an accuracy of 98.5% and an F1 score of 98.1%. Noviandy, Teuku Rizky; Idroes, Ghalieb Mutig; Maulana, Aga; Hardi, Irsan; Ringga, Edi Saputra; and Idroes, Rinaldi applied XGBoost and data augmentation techniques to a public credit card fraud dataset. The study achieved enhanced detection accuracy and reduced false positives, with an accuracy of 97.6% and an F1 score of 97.1%.Khalid, Abdul Rehman; Owoh, Nsikak; Uthmani, Omair; Ashawa, Moses; Osamor, Jude; and Adejoh, John proposed an ensemble approach that combines multiple machine learning algorithms using a public credit card fraud dataset. This approach improved overall detection performance, with an accuracy of 97.4% and an F1 score of 96.9%.

Cherif, Asma; Badhib, Arwa; Ammar, Heyfa; Alshehri, Suhair; Kalkatawi, Manal; and Imine, Abdessamad authored a systematic review of credit card fraud detection techniques in the context of disruptive technologies using various public datasets. The review offers a comprehensive overview of recent advancements and future directions without specific performance metrics.

III. DATASET DESCRIPTION

We utilized the Credit Card Fraud Detection Dataset 2023 from Kaggle, acquired using blockchain technology. This dataset consists of transactions made by European consumers throughout 2023, comprising approximately 550,000 anonymized records to protect user identities. The primary aim of this dataset is to assist in researching and developing fraud detection methods and models to identify potentially fraudulent transactions.

Key Features

- Each transaction has a unique identifier (id).
- V1-V28 represent anonymized features related to transaction parameters such as time and location.
- Amount: The transaction amount.
- Class: A binary label indicating whether a transaction is fraudulent (1) or not (0).

Potential Use Cases

- Develop machine learning methods to identify and prevent fraudulent credit card usage by detecting suspicious transactions based on specified features.
- Analyze the relationship between fraud and various merchant categories.
- Examine activity types to identify potential fraud risks.

Data Source: The data was compiled from credit card transactions by European consumers in 2023, with private data excluded to ensure confidentiality and comply with ethical standards.

IV. METHODOLOGY

4.1 Pre-processing and Feature Selection

In the initial stage of this research, we conducted a thorough data preprocessing phase to ensure the dataset's quality, integrity, and reliability. This step was crucial for developing accurate and effective fraud detection models. We began by carefully examining the dataset for duplicate entries, as duplicates could skew our analysis, lead to biased models, and compromise result validity. After identifying duplicates, we removed them to maintain a unique set of transactions, ensuring each data point represented a distinct event.

Next, we addressed missing values in the dataset, recognizing that such gaps could lead to inaccurate predictions, compromise model reliability, and undermine the trustworthiness of our findings. To mitigate these issues, we applied appropriate imputation techniques, selecting methods that suited the characteristics of our data and the needs of our machine learning algorithms. By filling in the missing values, we ensured that our dataset was complete, consistent, and ready for analysis. Following this, we converted the data types of each feature to the required formats, ensuring they matched the input expectations of our machine learning algorithms. This step was crucial to prevent errors, facilitate smooth processing during model training, and ensure accurate data interpretation by our algorithms.

Finally, we assessed the dataset for skewness, acknowledging that highly skewed features could adversely affect model performance, lead to biased predictions, and compromise the accuracy of fraud detection. By identifying and addressing skewness through suitable transformations and normalization techniques, we aimed to create a more balanced dataset. This approach was intended to help our machine learning algorithms learn patterns and relationships more effectively, ultimately improving the precision of fraudulent transaction detection.

4.2 Feature Engineering

We developed features to enhance model input, including:

- Time-based features: time since the last transaction, transaction frequency
- Behavioral features: spending patterns
- Transaction amount and location features

We analyzed feature correlations to identify strongly correlated and weakly correlated features, aiming to derive potential insights that could improve model performance.
The correlation analysis reveals a complex network of relationships among the parameters, with most showing high correlations with each other. This interdependence suggests that changes in one parameter may influence others. However, some parameters (V13, V15, V22, V23, V25, V26, and V28) show minimal correlation with many others, indicating they might be independent or have unique characteristics. Notably, the Amount parameter shows no correlation with any other parameter, suggesting it could be a dependent variable or outcome measure. In contrast, the Class parameter exhibits high correlations with several other parameters (V2, V3, V4, V9, V10, V11, V12, V14, and V16), highlighting its potential as a key factor influencing these parameters. Overall, these findings indicate that dimensionality reduction techniques could be effective in reducing the number of features while retaining crucial

information. Careful feature selection and analysis are essential to uncover meaningful relationships and patterns in the data.

4.3 Logistic Regression

Logistic regression is a statistical method used for binary classification. This predictive modeling technique estimates the probability of an outcome based on a set of independent variables. In logistic regression, the dependent variable is binary, meaning it can only take on two possible values, such as yes or no, true or false, or 0 or 1.

Here are some key characteristics of logistic regression:

- It is a linear model, meaning it assumes a linear relationship between the independent variables and the dependent variable.
- It employs a sigmoid function to convert the linear combination of the independent variables into a probability between 0 and 1.
- It is a widely used and easily interpretable machine learning model.

V. DECISION TREE

A decision tree is a machine learning model that classifies data using a tree-like structure. It includes internal nodes that represent tests on features, branches that denote the outcomes of these tests, and leaf nodes that signify the class labels. Here are some key characteristics of decision trees:

- They are easy to interpret and understand.
- They can handle both categorical and numerical data.
- They can be prone to overfitting if not properly pruned. Random Forest (RF) A Random Forest is an ensemble learning method that constructs multiple decision trees and outputs the class that is the mode for classification or the mean prediction for regression of the individual trees. It is known for its accuracy, robustness to overfitting, and ability to handle various data types.

Gradient Boosting Gradient boosting is an ensemble method that iteratively builds a model. Each new model is trained to correct the errors of the previous one. This process continues until a desired level of performance is achieved, often resulting in highly accurate models.LightGBM LightGBM is a gradient boosting framework known for its speed and accuracy. It uses tree-based learning algorithms and is optimized for large datasets. Key features include: • Gradient-based One-Side Sampling (GOSS) for faster training • Exclusive Feature Bundling (EFB) for handling categorical features • Leaf-wise tree growth for better accuracyXGBoost XGBoost is another gradient boosting framework optimized for speed and performance. It's widely used in machine learning competitions. Key features include: • Regularization to prevent overfitting • Support for various objective functions • Efficient implementation

RF + RNN (Hybrid Model) Combining a Random Forest (RF) with a Recurrent Neural Network (RNN) is an approach that leverages the strengths of both models. RFs are effective for capturing complex patterns in static data, while RNNs excel at handling sequential data. By combining them, it is possible to create models that can effectively handle both types of data. The specific architecture and training method would be tailored to utilize the strengths of both models. **Methodology** Developing such a hybrid model required careful consideration and experimentation. We made several customizations to the traditional architecture, incorporating attention mechanisms while maintaining a lightweight structure with 12,433 training parameters.

5.1 Implementation Ensemble Strategy:

The final fraud score is derived by combining the outputs of both RF and RNN models using a weighted average approach. Weights are determined by model performance metrics (e.g., precision, recall) on a validation set. Development was done using Python 3.12.1. Comprehensive descriptive data was generated with Python's pandas_profiling. Extreme cases in the data were identified using an outlier function, and the most relevant features were selected using scikit-learn's ExtraTreesClassifier. SMOTE was implemented with the imblearn module, setting k_neighbors to five. All models were built with the sklearn toolkit and validated using 5-fold cross-validation.

Logistic Regression (LR) was implemented with the LogisticRegression technique. A Decision Tree (DT) was developed using the DecisionTreeClassifier, with a maximum depth of four and 'entropy' as the criterion. SVM was developed with LinearSVC. The ensemble classifier used RandomForestClassifier with n_estimators set to 100. AdaBoost was modeled using AdaBoostClassifier with Gaussian kernels SVC for the base estimator, using default estimators and learning rates. XGBoost was modeled with XGBClassifier(), with hyperparameters tuned in subsequent steps.Initial findings indicate that RF is the most effective approach for detecting fraud using financial records from our dataset. RF's hyperparameters were tuned using GridSearchCV, improving its accuracy with 1,000 iterations and 5-fold cross-validation. The parameters tuned were:

- Learning rates (LR): [0.030, 0.010, 0.0030, 0.0010]
- Minimum child weights: [1.0, 3.0, 5.1, 7.2, 10.0]
- Gamma values: [0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 5.0]
- Subsamples: [0.60, 0.80, 1.00, 1.20, 1.40]
- Colsample_bytree values: [0.60, 0.80, 1.00, 1.20, 1.40]
- Max depths: [3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 12.0, 14.0]
- Regularization lambdas: [0.40, 0.60, 0.80, 1.0, 1.20, 1.40]

All parameter choices were tested, and the model was trained until the validation error (val_0-err) decreased across 10 folds.

VI. RESULTS AND DISCUSSION

The evaluated models demonstrated strong performance in credit card fraud detection, with accuracy, precision, recall, and F1-scores generally exceeding 95%. Despite all models showing promise, there were notable variations in performance and computational efficiency. Correlation analysis revealed a complex web of relationships between parameters, with many exhibiting high correlations with each other. This interdependence suggests that changes in one parameter can affect others. However, some parameters (V13, V15, V22, V23, V25, V26, and V28) showed almost no correlation with many others, indicating they might be independent or possess unique characteristics. Notably, the Amount parameter showed no correlation with any other parameter, hinting it could be a dependent variable or outcome measure. In contrast, the Class parameter exhibited high correlations with several others (V2, V3, V4, V9, V10, V11, V12, V14, and V16), suggesting it could be a key factor influencing these parameters. Overall, these findings suggest that dimensionality reduction techniques could effectively reduce the number of features while preserving important information, and that careful feature selection and analysis are necessary to uncover meaningful relationships and patterns in the data.

The RF+RNN model achieved the highest overall performance, excelling in identifying fraudulent transactions, though its training time was significantly longer compared to other models. Random Forest offered a balance of accuracy and efficiency, making it a viable option for many use cases. Both Gradient Boosting and XGBoost delivered robust results with relatively efficient training times. LightGBM provided a good compromise between accuracy and speed, making it suitable for resource-constrained environments. In contrast, Decision Tree and Logistic Regression models exhibited lower performance, particularly in identifying fraudulent transactions, resulting in low recall.

VII. CONCLUSION

This study presents a novel approach to improving credit card fraud detection by leveraging advanced machine learning techniques and blockchain technology. The integration of anomaly detection models, supervised learning methods such as Random Forest and Gradient Boosting, along with deep learning techniques like Recurrent Neural Networks (RNNs) and XGBoost, has demonstrated significant potential in identifying fraudulent activities. The results from the experimental analysis yielded impressive accuracy, with anomaly detection models achieving 99.9% accuracy, 99.8% recall, and an F1 score of 99.9%, underscoring the effectiveness of these approaches in detecting fraudulent transactions. The inclusion of blockchain enhances data security, traceability, and transparency, ensuring that detected fraud is verifiable and immutable. While these results are promising, further research is necessary to address real-world challenges such as scalability and the computational cost associated with processing large datasets. Additionally, the interpretability of complex deep learning models and their integration with blockchain require further investigation to ensure regulatory compliance and user trust. Overall, the combination of machine learning and blockchain holds great promise for revolutionizing credit card fraud detection, providing financial institutions with a more robust, secure, and efficient framework to combat evolving fraud tactics.

REFERENCES

1. Nguyen, T., & Hoang, D. T. (2020). A survey on credit card fraud detection using machine learning techniques. IEEE Access, 8, 188487-188504. https://doi.org/10.1109/ACCESS.2020.3028723

2. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797. https://doi.org/10.1109/TNNLS.2017.2736643
3. Zhou, Y., Wang, J., & Zhang, S. (2021). Blockchain and machine learning for cybersecurity and fraud detection in financial transactions. Future Generation Computer Systems, 115, 539-555. https://doi.org/10.1016/j.future.2020.09.016
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785
5. Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). Survey of fraud detection techniques. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(3), 125-139. https://doi.org/10.1109/TSMCC.2004.829513
6. Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences, 479, 448-455. https://doi.org/10.1016/j.ins.2019.01.069
7. Bahnsen, A. C., Aouada, D., Stojanovic, J., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142. https://doi.org/10.1016/j.eswa.2015.12.030
8. Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines. Proceedings of the International MultiConference of Engineers and Computer Scientists, 1, 442-447.
9. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings of the 2008 IEEE International Conference on Data Mining, 413-422. https://doi.org/10.1109/ICDM.2008.17
10. Nguyen, T. T., & Huynh, N. C. (2020). Blockchain application in financial services for fraud prevention: A review. Journal of Finance and Economics, 8(6), 263-271. https://doi.org/10.12691/jfe-8-6-1
11. Jurgovsky, J., Granitzer, M., Ziegler, K., & Calabretto, S. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234-245. https://doi.org/10.1016/j.eswa.2018.01.037
12. Xiao, Y., Xing, C., & Zhao, X. (2019). A survey on anomaly detection for credit card fraud recognition. Procedia Computer Science, 162, 389-397. https://doi.org/10.1016/j.procs.2019.12.024

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com